Music Recommendation System Driven by Facial Expression Recognition

Taro Asada[1], Motoki Kawamura[2], Yasunari Yoshitomi[1], Masayoshi Tabuse[1]
1: Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: t_asada@mei.kpu.ac.jp, {yoshitomi, tabuse}@kpu.ac.jp
http://www2.kpu.ac.jp/ningen/infsys/English_index.html
2: FUJITSU LIMITED
Shiodome City Center, Higashi Shimbashi 1-5-2,Minato-ku, Tokyo 105-7108, Japan

Abstract

This paper reports on modifications to our previously proposed music recommendation system to include a method by which users interact with the system via facial expressions. More specifically, by presenting either happy or neutral facial expressions to a personified agent, the user informs the system of his or her opinion regarding the song being played. A happy facial expression means that he or she would enjoy listening to the song again, while a neutral facial expression means the opposite.

Keywords: Music recommendation system, Music therapy, Facial expression synthesis, MMDAgent, Microsoft Face application programming interface

1. Introduction

Recently, music therapy has been used to improve the recognition ability of people, particularly older people. Music therapy may be more effective if music that is liked by an individual is adopted.

In the system proposed in our previous study,[1] it is necessary for subjects to enter subjective information into the computer using a keyboard to report their evaluations to the system, which would then propose music recommendations. However, keyboard user input is often a daunting task for physically disabled or elderly persons (especially those with dementia), which means that support from nurses, caregivers such as family or facility staff, etc., is required.

With that point in mind, this paper reports on improvements to that system made by including an interface through which users interact with the system via a personified agent. More specifically, our improved system recognizes and analyses facial expressions and uses synthesized voice/expression output when producing recommendations. We also report on a performance evaluation of our improved system.

2. Face API

In our newly proposed system, facial expressions in the input image are detected using the Face application program interface (API),[2] which has already been used in several previous studies.[3,4] Face, which is a Web API provided by Microsoft as part of its artificial intelligence (AI) related Cognitive Services, is designed to identify eight emotional states: anger, contempt, disgust, fear, happiness, a neutral state, sadness, and surprise. An example result obtained using the Face API is shown in Fig. 1.

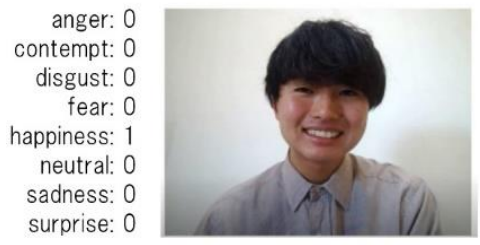

Fig. 1. Facial expression recognition using Face API (smile).

3. Music recommendation system based on facial expression recognition results

3.1. *Improvements over the previously proposed system*

Our previously proposed system[1] required keyboard inputs in order to play music, select the genre of the recommended music, input the user's subjective evaluation of the recommended music, or to play the next piece of music.

In contrast, we implemented two improvements in our proposed music recommendation system to accommodate subjects who have difficulty in inputting information via keyboards so they can use it unassisted:

(i) We prepared video footage of a personified agent[5] giving music recommendations by the method outlined in the previous study[6] and implemented it in the new system.

(ii) We modified the system's genre recommendation method by using facial expressions as system input for obtaining users' subjective evaluations. More specifically, user facial expressions are obtained in response to each piece of music presented by the system and are analyzed using the Face API to produce recommendations.

With these improvements, system users can operate the improved music recommendation system independently without having to manipulate a keyboard.

3.2. *System overview*

Initially, the subjects are asked to select one of the recommended song genres, either children's songs or popular music, by consciously making facial expressions to display feelings such as "happy" for the children's songs or "neutral" for the popular selections. Then, songs from the selected music genre are played by the music recommendation system as described in our previous report.[1] The subjects are instructed to consciously present "happy" or "neutral" facial expressions when they felt enjoyment (would want to listen to that piece again) or when they felt no enjoyment (would not want to listen to that piece anymore), respectively.

The music recommendation system then selects and plays the next recommended song while considering the answer expressed by the user's facial expression as a subjective evaluation value. If the total number of recommended songs reaches the system's upper limit for the recommendations, or if there are no more songs available to evaluate, the music recommendation process is finished, as shown in Fig. 2.

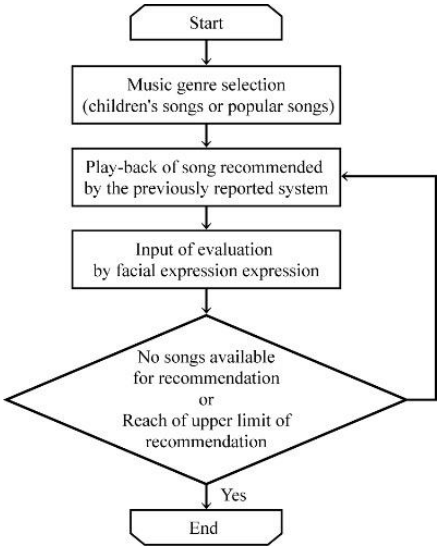

Fig. 2. Processing flow of the proposed system.

3.3. *Input using facial expression recognition*

As discussed above, in our previous system[1], the subjective evaluation value of the user was input using the keyboard to determine the next recommended song. In this study, the system was modified so that the "I would enjoy listening to this again (1)" is input when the happiness value obtained from Face API exceeds the threshold value, while in other cases, "I do not want to listen to this again (0)" is obtained.

4. Experiment

4.1. *Preliminary experiment*

4.1.1. Setting the threshold

In order to determine the best way to use facial expressions as evaluation inputs for our music recommendation system, it was first necessary to conduct a preliminary experiment to define the threshold value.

4.1.2. Conditions

The music recommendation system developed in this study was used on 5 male subjects in their 20s. The experiment was performed using two databases: one

consisting of 52 children's songs and another consisting of 58 popular songs.

4.1.3. Results

The threshold was determined to be 0.5, which is the average of the maximum and minimum values of "happiness" for the five subjects, obtained from the preliminary experiment.

4.2. *Evaluation experiment*

4.2.1. Conditions

In this experiment, 10 test subjects (7 men and 3 women in their 20s) were first asked to obtain music recommendations using the previously reported music recommendation system[1]. Next, the subjects were asked to obtain music recommendations using the improved system proposed in this study, after which they answered a questionnaire consisting of five-grade evaluations for three questions. In addition, the subjects were tasked with submitting their individual subjective evaluation values for the recommended music. This experiment was performed using the same two databases mentioned in Section 4.1.2. above.

4.2.2. Results and discussion

Table 1 shows the contents of the questionnaire items, while Table 2 shows the evaluation values on each question item for all subjects, and Table 3 shows the average value of the items in each question. Question items 1 and 2 addressed the operability and usability of the music recommendation system developed by our method, while Question item 3 was answered by comparing the previous version of our system[1] with the newly proposed version.

Table 4 shows the concordance rate of facial expression recognition and subjective evaluation, and recommendation accuracy for each subject, while Table 5 shows the concordance rate of facial expression recognition and subjective evaluation, and recommendation accuracy.

These results show that our proposed system has a good level of usability and a high concordance rate of facial expression recognition and subjective evaluation, which means that it is possible to input user evaluations into the music recommendation system using facial expressions rather than keyboards.

However, since there were two subjects who answered that "the music playback time was too long" in the free description, it will be necessary to include a feature that will allow the subject to interrupt the music playback and advance to the next music selection at will.

Table 1. Questionnaire used to evaluate the proposed system.

No.	Question
1	How easy was it to express your subjective evaluation via facial expressions? [5] Very easy, [4] Easy, [3] Neither easy nor difficult, [2] Difficult, [1] Very difficult
2	Was the agent's description easy to understand? [5] Very easy, [4] Easy, [3] Neither easy nor difficult, [2] Difficult, [1] Very difficult
3	Did you find the proposed music recommendation process to be more fun to use than the previous system? [5] Extremely fun, [4] Fun, [3] Neither, [2] Not too fun, [1] Not fun at all

Table 2. Evaluation of the proposed system

Question no.	Subjects									
	F	G	H	I	J	K	L	M	N	O
.1	3	5	5	5	5	5	5	5	5	5
.2	4	4	4	5	5	5	5	4	5	5
.3	5	4	5	4	4	5	5	4	5	5

Table 3. Average value of evaluation of the proposed system

Question no.	1	2	3
Average	4.8	4.6	4.6

Table 4. Concordance rate of ①facial expression recognition and ②subjective evaluation, and recommendation accuracy for each subject

Subjects	F	G	H	I	J
Genre	Children's song	Popular song	Children's song	Children's song	Children's song
Concordance rate of ① and ② (%)	100	100	100	100	100
Recommendation accuracy (%)	57.1	66.7	64.3	50.0	50.0

Subjects	K	L	M	N	O
Genre	Children's song	Children's song	Popular song	Popular song	Popular song
Concordance rate of ① and ② (%)	93.3	100	100	100	100
Recommendation accuracy (%)	78.6	50.0	55.6	50.0	77.8

Table 5. Concordance rate of facial expression recognition and subjective evaluation, and recommendation accuracy

Concordance rate (%)	99.2
Recommendation accuracy (%)	59.7

Also, in the free comments, one subject noted, "If I am not looking at the camera when the system makes a determination as to whether 'I don't want to listen to it again,' my expression is judged to be neutral even if I am showing a different expression."

This situation occurs because our system only makes an "I would enjoy listening to this again (1)" determination when the happiness value obtained from the Face API exceeds the threshold value and makes an "I do not want to listen to this again (0)" determination in all other cases. Therefore, it will be necessary to use a more precise method for neutral facial expression determinations.

5. Conclusion

In this study, we demonstrated the effectiveness of using conscious facial expressions as music evaluation inputs in a music recommendation system. However, to achieve further usability, improvements such as allowing the user to interrupt the playback of a song and start the playback of the next song selection at will should be considered.

Acknowledgments

We would also like to thank the subjects of our experiments for their cooperation.

References

1. Y. Yoshitomi, T. Asada, R. Kato, Y. Yoshimitsu, M. Tabuse, N. Kuwahara, and J. Narumoto, “Music Recommendation System through Internet for Improving Recognition Ability Using Collaborative Filtering and Impression Words”, Journal of Robotics, Networking and Artificial Life, Vol.2, No.1, pp. 54-59, 2015.
2. Microsoft Azur, “Face API”, https://azure.microsoft.com/ja-jp/services/cognitive-services/face/
3. K. Kumazaki, S. Shiramatsu, “Considering Method for Estimating Atmosphere of Debate using RealSense camera”, Proceedings of the 80th National Convention of IPSJ, pp. 287-288, March, 2018. (In Japanese)
4. A. Nakamura, T. Takizawa, T. Hoshi, H. Tsunashima and Q. Chen, “Automatic Font Generation Algorithm Based on Image Kansei”, Journal of Japan Society of Kansei Engineering, Vol. 17, No. 5, pp. 523-529, October 2018. (In Japanese)
5. A. Matsui, M. Sakurai, T. Asada, Y. Yoshitomi, and M. Tabuse, “Music Recommendation System Driven by Interaction between User and Personified Agent Using Speech Recognition, Synthesized Voice and Facial Expression”, Proceedings of International Conference on Artificial Life and Robotics, pp.28-31, 2021.
6. T. Asada, R. Adachi, S. Takada, Y. Yoshitomi, and M. Tabuse, “Facial Expression Synthesis System Using Speech Synthesis and Vowel Recognition”, Journal of Advances in Artificial Life Robotics, Vol. 1, No. 2, pp. 59-63, 2020.

Authors Introduction

Dr. Taro Asada

He received his B.S., M.S. and Ph.D. degrees from Kyoto Prefectural University in 2002, 2004 and 2010, respectively. He works as an Associate Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are human interface and image processing. HIS, IIEEJ member

Mr. Motoki Kawamura

He received his B.S. degree from Kyoto Prefectural University in 2021. He works at FUJITSU LIMITED.

Mr. Yasunari Yoshitomi

He received his B.E, M.E. and Ph.D. degrees from Kyoto University in 1980, 1982 and 1991, respectively. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His specialties are applied mathematics and physics, informatics environment, intelligent informatics. IEEE, HIS, ORSJ, IPSJ, IEICE, SSJ, JMTA and IIEEJ member.

Dr. Masayoshi Tabuse

He received his M.S. and Ph.D. degrees from Kobe University in 1985 and 1988 respectively. From June 1992 to March 2003, he had worked in Miyazaki University. Since April 2003, he has been in Kyoto Prefectural University. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are machine learning, computer vision and natural language processing. IPSJ and IEICE member.